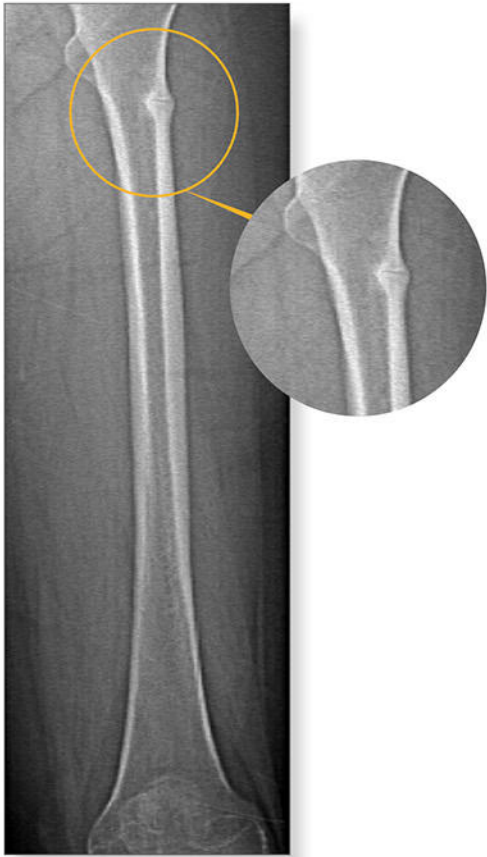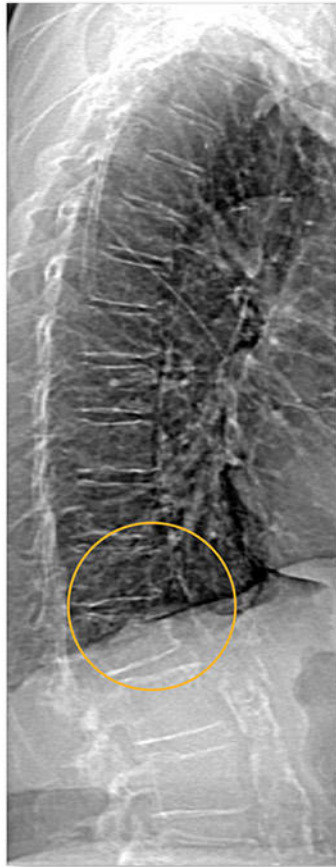![HOLOGIC® The Science of Sure]

Powerful images.
Clear answers.

Manage Patient's concerns about Atypical Femur Fracture*

Vertebral Fracture Assessment – a critical part of a complete fracture risk assessment

Advanced Body Composition® Assessment – the power to see what's inside

Contact your Hologic rep today at insidesales@hologic.com

www.hologic.com | **info@hologic.com** | **1.781.999.7300**

The Emperor's New Clothes: What Randomized Controlled Trials Don't Cover

John A Eisman,[1] P Geusens,[2] and J van den Bergh[3]

[1]Bone Biology Division, Garvan Institute of Medical Research, Darlinghurst, Sydney, NSW, Australia; School of Medicine Sydney, University of Notre Dame Australia, Darlinghurst, Sydney, NSW, Australia; St Vincent's Clinical School, Faculty of Medicine, UNSW Sydney, Sydney, NSW, Australia; Endocrinology, St Vincent's Hospital, Sydney, NSW, Australia; Care and Public Health Research Institute, Maastricht University Medical Center, Maastricht, The Netherlands
[2]Department of Internal Medicine, VieCuri Medical Centre, Venlo, The Netherlands; Department of Internal Medicine, Maastricht University Medical Centre, Maastricht, The Netherlands; Faculty of Mediciney, University Hasselt, Maastricht, Belgium
[3]Department of Rheumatology, Maastricht University, The Netherlands; Faculty of Medicine, University Hasselt, Belgium

Introduction

Randomized controlled trials (RCTs) are the bedrock of evidence-based medicine (EBM). They avoid the major pitfalls of anecdotal evidence and conscious as well as unconscious bias in patient selection and outcome assessments. RCTs have greatly enriched our knowledge of the efficacy and safety of medicines across all aspects of health care, including surgical interventions.

As a prerequisite for publication, most major international journals require that RCTs are registered with one or more government-sponsored database (eg, clinicaltrials.gov, a resource provided by the U.S. National Library of Medicine in Bethesda, MD) prior to the enrollment of the first patient. This simple condition minimizes the potential "cherry-picking" of studies that show a positive finding. Although there is no requirement that ensures or mandates that all studies be published, this condition is designed to prevent the selective reporting and publication of research outcomes and the suppression of less-favorable outcomes.

Two critical aspects of RCTs are the random allocation of participants to a drug being tested versus placebo (or another drug) and the blinded assessment of outcomes for both primary efficacy and adverse events. These aspects avoid an unbalanced allocation of subjects to the test agent and, most importantly, a biased assessment of outcomes—positive or negative—to the test agent or placebo. RCTs, as noted above, have revolutionized the assessment of benefit and risk for new therapeutic interventions.

However, although RCTs have become the bedrock of clinical science, there are many aspects that RCTs are not designed to cover. In this Perspective, we highlight some important gaps in the coverage of RCTs and describe alternative settings where they can be offset. RCTs leave major gaps, and as in the *Emperor's New Clothes*, these gaps are largely ignored by authorities (such as experts and drug-approval agencies) who should recognize them explicitly and try to address them. This Perspective will address: (1) study entry unrepresentativeness; (2) the related communication issue of the misleading estimates of number-needed-to-treat (NNT); and (3) how the statistical limit of a primary study outcome is not applied to adverse events.

Gaps in Randomized Controlled Trials

The first major gap is that RCT study entry criteria are largely unrepresentative of likely future community-based patients. This is a recognized issue when drugs shown to be effective in high-fracture-risk groups are inappropriately recommended for or used in lower-fracture-risk individuals. This leads to a second gap: the misleading estimates of the NNT when translating the RCT findings to real-world individuals. These NNTs are "optimistic" in a low-risk group, and "pessimistic" in a high-risk group. NNT is an illogical concept unless the underlying risk is known.[1]

The third major gap is that the critical statistical logic of a primary study outcome is not applied to adverse events.

Given these gaps, there is a critical need to complement RCT data with real-world cohort studies adjusted for known biases to provide more robust information on risks and benefits of treatments. It is worth noting that missing data are a challenge in any study, whether an RCT or a cohort-based study. Careful assessment of those entering a study and those assessed at the end of a study needs to be explicitly stated in both approaches.

Study entry unrepresentativeness

Pivotal RCTs of chronic health conditions minimize confounding by excluding other, often common, health conditions. These criteria, excluding 80% to 90% of future patients,[2,3] result in "super healthy" cohorts, unrepresentative of future patients.

Interestingly, although placebo mortality was around 1% per annum in most of these RCTs, in the one with very open-entry criteria posthip fracture, it was approximately 7% per annum.[4] Although older patients posthip fracture may be expected to have a higher mortality rate, this sevenfold difference is consistent with RCTs enrolling uncommonly healthy individuals. In longer, larger studies for efficacy and safety endpoints, study cohorts are becoming progressively less representative. In a meta-analysis of pivotal RCTs of anti-osteoporosis agents in osteoporosis, there was approximately a 10% overall benefit on survival.[5] However, the finding of this meta-analysis of RCTs has not found its way into treatment guidelines. Thus, pivotal RCTs may not be studying real future patients. In community practice, the efficacy of specific desired outcomes (and some other important outcomes) in the RCT research setting may be substantially different from the effectiveness of the same outcomes in the real-world situation.

Misleading number-needed-to-treat

A related issue with RCT reporting is the common focus on NNT: estimated as individuals needed to be treated for a benefit. These numbers depend on the underlying absolute risk and on the duration of treatment that is seldom mentioned explicitly. Thus, two treatments with identical efficacy would have substantially different NNTs if one study's NNT were calculated for 2 years and the other for 3 years. Even if this misleading time-dependent presentation were overcome, underlying risks largely drive NNT. Hence, the same treatment would have a much lower NNT if the study was conducted in a higher-risk group than in a lower-risk group. Thus, based on simple mathematics, the NNT for efficacious agents is largely determined by absolute risk in the placebo arm.[1] Number needed to harm, often identified in postmarketing surveillance, suffers from even greater uncertainty about exposure, duration, and underlying risks. Both, if used at all, should be expressed relative to exposure duration and underlying risk.

Statistical limit to one primary study outcome not applied to adverse events

Limiting evaluated outcomes in clinical trials to a single prespecified primary outcome is the statistical bedrock of EBM.[6] However, the same logic is not applied to adverse effects, each of which is usually considered as though it were a prespecified primary outcome, ie, without adjustment for multiple testing. For example, in the RCT of zoledronic acid in individuals following a hip fracture,[4] mortality (not a primary outcome) was 28% lower in the treated arm. This biologically important, but not prespecified outcome is discounted. Had there been a smaller but "significant" increase in mortality, we suggest the drug would no longer be available. This perverse logic plagues progressively larger RCTs, particularly for unexpected adverse outcomes: These should be subject to the same rigorous evaluation in two or more studies, in case chance adverse findings are considered real.

The suggestion to lower the p-value for statistical significance to 0.005[7] could reduce false-positives for specific study outcomes, but would markedly increase RCT study size and could make many RCTs unfeasible. The current approach of retaining the $p < 0.05$ threshold, but requiring similar specific outcomes in two similar independent studies has the effect of lowering the threshold for a consistent "positive" outcome equivalent to $p < 0.0025$. This approach has the benefit of

increasing the certainty around outcomes, while retaining studies of manageable size and duration.

Whatever approach is taken, critically it must be applied to adverse events as well as primary outcomes. Unless there is a rationale for a specific adverse outcome based on known treatment mechanisms, the significance threshold should be formally adjusted for the number of potential adverse outcomes examined. We suggest unexpected adverse events should be viewed cautiously rather than overly defensively, as is usually the case.

Limitations of RCTs compared to real-world studies

Examples of these limitations of RCTs compared with real-world evidence are available in many fields of medicine such as oncology,[8] but also in the field of fracture prevention studies.

Although in one RCT, zoledronic acid treatment was associated with an increased risk of atrial fibrillation, meta-analyses of all bisphosphonates (BPs) showed discordant results; nevertheless, two large health care plans' databases showed no association.[9]

On the other hand, after a recent hip fracture, zoledronic acid treatment decreased the risk of mortality.[4] Meta-analyses of BPs indicate a decreased risk of mortality, albeit with modest effect (~10%) in the unrepresentatively healthy RCT subjects.[5] Real-world cohorts have indicated a larger effect of BPs on mortality.[10–13] However, despite the beneficial effects on survival achieved with zoledronic acid in a posthip fracture RCT and meta-analyses of RCTs, as well as confirmation in observational studies, there has been no consensus. This is not consistent with applying a single logical standard for beneficial and adverse effects.

Although RCTs with BPs and denosumab individual cases of atypical femur fractures (AFFs) and osteonecrosis of the jaw (ONJ) have been reported, observational cohort studies identified the association with an increased relative risk in patients on BPs or denosumab, but with a low absolute risk and thus low statistical power in RCTs.[14,15] In addition, it was reported from observational studies that lower limb geometry and Asian ethnicity[16] may contribute to the risk of an AFF. These studies also provided a better definition of these adverse effects. The observational studies regarding an AFF and ONJ identified the adverse effect and refuted the RCT findings for atrial fibrillation. Based on these studies, an international consensus was reached.

By addressing the real world, cohort and database studies can overcome some of these limitations provided their own biases are recognized and addressed. By examining data from a large body of patients receiving treatment for longer periods, both adverse events and unexpected benefits can be appreciated, provided the denominator of numbers treated and duration of treatment and follow-up are robust.

Critically, cohort and database studies can have oversight of adverse events across much larger numbers of people, of different health status, and over much longer periods. An outcome, such as mortality, though usually collected in RCTs, can be followed in cohort and database studies for much larger groups and over much longer periods. It is unlikely—for ethical reasons—for mortality to be the primary outcome measure in any RCT of individuals with fractures, as such a trial would be withholding recommended therapy to individuals with the explicit expectation that they would risk earlier mortality.

Clothing the Emperor

On the beneficial effects on survival achieved with zoledronic acid in a posthip fracture RCT in meta-analyses of RCTs, and confirmation in observational studies, there has been no consensus. This is not consistent with applying a single logical standard for beneficial and adverse effects.

Postmarketing databases and cohorts provide critical information, highly complementary to pivotal RCTs, albeit with their own biases. As noted below, depending on the quality of data available, each of these biases can be addressed. Cohort studies have the advantage of more detailed data collection, whereas population databases have the advantage of both time and size. Missing data, less-structured predictor and outcome assessment, and unknown residual confounding are considerations, albeit largely absent from RCTs, that can affect cohort and databases studies. However, these could be considered as random rather than specific to treatment, except for three major biases:

1. Patient selection: Treatment offered to healthier or to sicker patients distorts outcomes; hence, these criteria need to be specifically addressed. This is sometimes noted as confounding by indication. If an efficacious drug is restricted to those at highest risk, those treated may still have an adverse outcome, but better than might be predicted based on those underlying characteristics. This bias can be partially addressed if the outcome predictors can be assessed.
2. Immortal time bias: Survival until a particular therapy is available inevitably selects healthier individuals. Cohort or national medication possession databases can narrow this gap by limiting comparisons between matched treated and untreated individuals alive when the treatment is recommended or initiated.
3. Healthy complier bias: Healthier people are more likely to seek and adhere to treatments, such as compliers to placebo therapy.[17] This bias can be mitigated in cohort studies, as in RCTs, by using a treatment recommended (equivalent to intention to treat) rather than a treatment used (equivalent to per protocol) analysis.

Cohort and database studies, given awareness of these biases and applying tools to address and overcome them, can provide invaluable complementary data that may never be achievable through RCT approaches.

In summary, despite the critical importance of RCTs to advance medical science, they still leave some disconcerting holes in our knowledge about efficacy and safety. There is a clear benefit to complementing RCT data with cohort and database studies adjusted for potential biases. Applying Bayesian approaches (ie, most likely true values rather than simplistic *p*-values) to Cochrane data collections[7,18–20] may provide more complete and robust information on relative risks and benefits of available treatments. The Emperor of EBM is sometimes rather thinly covered, but there are ways to clothe this nakedness.

Disclosure

The authors state they have no competing interests within the scope of this paper.

References

1. Seeman E, Eisman JA. 7: Treatment of osteoporosis: why, whom, when and how to treat. The single most important consideration is the individual's absolute risk of fracture. Med J Aust 2004; 180:298–303.
2. Reyes C, Pottegard A, Schwarz P, et al. Real-Life and RCT Participants: alendronate users versus FITs' trial eligibility criterion. Calcif Tissue Int 2016;99:243–9.
3. Bernhardt J, Raffelt A, Churilov L, et al. Exploring threats to generalisability in a large international rehabilitation trial (AVERT). BMJ Open 2015;5(8):e008378.
4. Lyles KW, Colon-Emeric CS, Magaziner JS, et al. Zoledronic acid and clinical fractures and mortality after hip fracture. N Engl J Med 2007;357:1799–809.
5. Bolland MJ, Grey AB, Gamble GD, et al. Effect of osteoporosis treatment on mortality: a meta-analysis. J Clin Endocrinol Metab 2010;95:1174–81.
6. Halsey LG, Curran-Everett D, Vowler SL, et al. The fickle P value generates irreproducible results. Nature Meth 2015;12:179–85.
7. Ioannidis JPA. The proposal to lower P value thresholds to .005. JAMA 2018:E1–E2. doi:10.1001/jama.2018.1536
8. Schilsky RL. Finding the evidence in real-world evidence: moving from data to information to knowledge. J Am Coll Surg 2017; 224:1–7.
9. Bunch TJ, Anderson JL, May HT, et al. Relation of bisphosphonate therapies and risk of developing atrial fibrillation. Am J Cardiol 2009;103:824–8.
10. Brozek W, Reichardt B, Zwerina J, et al. Antiresorptive therapy and risk of mortality and refracture in osteoporosis-related hip fracture: a nationwide study. Osteoporos Int 2016;27:387–96.
11. Bondo L, Eiken P, Abrahamsen B. Analysis of the association between bisphosphonate treatment survival in Danish hip fracture patients— a nationwide register-based open cohort study. Osteoporos Int 2013;24:245–52.
12. Center JR, Bliuc D, Nguyen ND, et al. Osteoporosis medication and reduced mortality risk in elderly women and men. J Clin Endocrinol Metab 2011;96:1006–14.
13. van Geel T, Bliuc D, Geusens PPM, et al. Reduced mortality and subsequent fracture risk associated with oral bisphosphonate recommendation in a fracture liaison service setting: a prospective cohort study. PLoS One 2018;13(6):e0198006. doi:10.1371/journal. pone.0198006
14. Shane E, Burr D, Abrahamsen B, et al. Atypical subtrochanteric and diaphyseal femoral fractures: second report of a task force of the American Society for Bone and Mineral Research. J Bone Miner Res 2014;29:1–23.
15. Khan AA, Morrison A, Hanley DA, et al. Diagnosis and management of osteonecrosis of the jaw: a systematic review and international consensus. J Bone Miner Res 2015;30:3–23.
16. Dell RM, Adams AL, Greene DF, et al. Incidence of atypical nontraumatic diaphyseal fractures of the femur. J Bone Miner Res 2012;27:2544–50.
17. Stetler C. Adherence, expectations and the placebo response: why is good adherence to an inert treatment beneficial? Psychol Health 2014;29:127–40.
18. Nguyen ND, Eisman JA, Nguyen TV. Anti-hip fracture efficacy of bisphosphonates: a Bayesian analysis of clinical trials. J Bone Miner Res 2006;21:340–9.
19. van Geel TA, Nguyen ND, Geusens PP, et al. Development of a simple prognostic nomogram for individualising 5-year and 10-year absolute risks of fracture: a population-based prospective study among postmenopausal women. Ann Rheum Dis 2011;70:92–7.
20. Zhang L, Pang Y, Shi Y, et al. Indirect comparison of teriparatide, denosumab, and oral bisphosphonates for the prevention of vertebral and nonvertebral fractures in postmenopausal women with osteoporosis. Menopause 2015;22:1021–5.